

Accuracy and precision of dolphin group size estimates

Tim Gerrodette^{1,*}, Wayne L. Perryman¹, Cornelia S. Oedekoven²

¹ NOAA Fisheries, Southwest Fisheries Science Center, 8901 La Jolla Shores Dr., La Jolla, California 92037, USA

² Centre for Research into Ecological and Environmental Modelling, The Observatory, Buchanan Gardens, University of St. Andrews, St. Andrews, Fife KY169LZ, Scotland, UK

*Corresponding author: tim.gerrodette@noaa.gov

Abstract: Estimating the number of dolphins in a group is a challenging task. To assess the accuracy and precision of dolphin group size estimates, observer estimates were compared to counts from large-format vertical aerial photographs. During 11 research cruises, a total of 2,435 size estimates of 434 groups were made by 59 observers. Observer estimates were modeled as a function of the photo count in a hierarchical Bayesian framework. Accuracy varied widely among observers, and somewhat less widely among dolphin species. Most observers tended to underestimate, and the tendency increased with group size. Groups of 25, 50, 100, and 500 were underestimated by <1%, 16%, 27%, and 47%, respectively, on average. Precision of group size estimates was low, and estimates were highly variable among observers for the same group. Predicted true group size, given an observer estimate, was larger than the observer estimate for groups of more than about 25 dolphins. Predicted group size had low precision, with coefficients of variation ranging from 0.7 to 1.9. Studies which depend on group size estimates will be improved if the tendency to underestimate group size and the high uncertainty of group size estimates are included in the analysis.

Keywords: group size estimation, abundance estimation, aerial photography, Bayesian hierarchical model, random-effects model, reverse jump MCMC

Estimation of group size is an important component of ecological and behavioral studies of animals which occur in groups. However, estimation of group size in wildlife studies can be difficult. Replicate counts of birds showed high variation (Ryan and Cooper 1989), the number of birds was undercounted in aerial surveys (Bayliss and Yeomans 1990), and known group sizes of elk were underestimated from a helicopter (Cogan and Diefenbach 1998). Even counting the number of birds in photographs had a negative bias (Erwin 1982). Experiments in visual perception have shown a tendency to underestimate the size of large groups of objects (Krueger 1972), apparently related to distortions produced by saccadic (“jerky”) eye movements (Binda *et al.* 2011). Determining the size of a group of cetaceans is particularly challenging because of several characteristics that make group size estimation difficult: (1) the animals are moving; (2) an unknown fraction of the group is underwater at any moment; (3) the fraction underwater changes with behavior; (4) groups can be large; and (5) the distribution of group sizes is usually skewed, with a few groups much larger than the mean.

Accurate estimation of group size is necessary for unbiased estimation of abundance. In standard distance sampling (*e.g.*, line transects), the density of groups is estimated and then multiplied by an estimate of expected group size (Buckland *et al.* 2001). Alternatively, group size may be a covariate of the detection process and expected group size is not estimated explicitly (Borchers and Burnham 2004). In either case it is assumed that group sizes are measured accurately. Using earlier subsets of the photographic calibration data presented here, some line-transect analyses have used group size estimates corrected by observer-specific calibration factors (Gerrodette and Forcada 2005, Barlow and Forney 2007). In most studies, however, correction factors for group size estimation are not available.

Assessing precision of group size estimates is equally important. Even if group sizes were to be estimated accurately on average, there is measurement error associated with each group size estimate. Including the variability associated with group size estimates is necessary for proper assessment of uncertainty. If measurement error is not included, variance of estimates of abundance and other quantities that depend on group size estimates will be too small. In other words, an important source of uncertainty will not be included in the analysis, and conclusions may appear to be more precise than they should be.

In this large field study, we measured the accuracy and precision of dolphin group size estimates. True group size was assessed with counts from high-quality vertical aerial photographs, and ship-based observer estimates were calibrated against these counts. The tendencies of different individual observers to under- or over-estimate group size were estimated in a hierarchical Bayesian framework, for different group sizes, species, and sea-state conditions. The performance of a new (out-of-sample) observer was predicted by integrating over observer and/or species effects. Given an observer estimate, we inferred true group size by sampling posterior distributions.

Methods

Field methods

Photographs of dolphin groups were collected during 11 research cruises between 1987 and 2006 in the eastern tropical Pacific Ocean. During all cruises except the last, the NOAA vessel *David Starr Jordan* carried a Hughes 500D helicopter equipped with two large-format military reconnaissance cameras mounted below the fuselage. During the 2006 cruise, images

were collected with the same camera systems mounted in a NOAA Twin Otter fixed wing aircraft. Under conditions of sun angle (generally mid-morning and mid-afternoon) and sea state (generally Beaufort 0-4) that allowed dolphins to be clearly visible from above, vertical photographs of dolphin groups were taken from an altitude of 200-300m (Gilpatrick 1993). The camera recorded images on 114mm negatives, and had a motion-compensation system that moved the film at the same speed that the image was moving within the camera, thus eliminating blurring due to the forward motion of the aircraft. The cycle rate of the camera was adjusted to achieve 80% overlap between adjacent frames during a photographic pass over a dolphin group. The number of photographic passes of each dolphin group varied with group size, configuration and behavior.

After a group of dolphins had been photographed, the group was approached by the ship in a way to give the marine mammal observers on the ship the best possible view of the whole group, considering wind, swell, and sun angle. All observers who had adequate views of the group, usually all six observers on the ship, made their best estimates of group size. We refer to these estimates as the “observer estimates.” Observers usually first detected dolphins with 25X binoculars, but switched to 7X binoculars and then to naked eye as the ship approached the group. The minimum approach distance varied with group size and behavior, but typically was 10-50m. Observers made group size estimates independently and did not discuss their estimates with each other, either during the sighting or afterward. Independence in this context refers to the behavioral independence of the observers, not to the statistical independence of their estimates. All observers had previous experience in cetacean field work. Before each cruise, observers were given training on group size estimation, including tests with known numbers of

static objects, computer simulations of moving, intermittently visible objects, and instruction on counting by subgroups (*e.g.*, by tens or fifties) for more consistent estimation.

Laboratory methods

The aerial photographs of dolphin groups were reviewed on light tables equipped with dissection microscopes (Gilpatrick 1993). Photographs were compared with notes recorded during the photographic passes to ensure that the entire group was captured within the series of images that made up a photograph pass. For groups that were successfully photographed, the best pass was selected, and three readers independently counted the number of dolphins in the group from the series of images. If the CV among counts was > 0.1 , or if notes by aerial and shipboard observers indicated that there was confusion over the identity of the group, the group was not included in the data analyzed here (Gilpatrick 1993).

To qualify as a “calibration school” for this analysis, the whole group had to be photographed from the air with a series of overlapping photographs, the photo counts of the three independent readers had to agree closely, and the shipboard observers had to view the whole group for a sufficient time to make good estimates. Calibration schools were thus not a random sample of all dolphin groups, but rather a selected set for which we were confident that true group size could be accurately determined. We omitted as outliers eight cases for which there was a large (greater than a factor of four) discrepancy between mean photo count and mean observer estimate, probably a result of undetected splitting or coalescence of groups after photography but before observer estimates. A total of 434 groups met these criteria as calibration schools, with 2,435 estimates of group size by 59 observers.

Statistical model

To evaluate observer estimates of group size, we used the mean of the counts by the three photograph readers for each calibration school, and refer to this as the “photo count.” This measure of true group size had some error (variation among the three readers), but this variation (mean photo count CV over all groups = 0.047) was much smaller than the variation among observer estimates of the same groups (mean CV = 0.42). Preliminary exploration of the data suggested that, on a log-log scale, observer estimates could be linearly related to photo counts and that variance was approximately constant over a large range of group sizes (Fig. 1A). In addition, observers varied widely in the accuracy of their group size estimates (Fig. 1B). We evaluated a variety of linear and nonlinear models in a frequentist setting, with both fixed and random effects, with R function *lmer*, and used likelihood ratio tests, information criteria such as AIC and DIC, and visual examinations of residual and q-q plots to identify a reasonable set of candidate models. We found that dolphin species and Beaufort sea state could possibly affect the accuracy of group size estimates, and that a linear model of the logarithm of photo counts provided a more parsimonious fit to the data than a quadratic model.

Let y_{ij} be the observer estimate of the size of group i by observer j , and let x_i be the photo count of group i . We modeled differences among observers as random effects, and dolphin species and wind conditions as fixed additional effects that might affect group size estimates.

The full hierarchical model may be written as

$$\begin{aligned}
 \log(y_{ij}) &= \alpha_{0j} + (\alpha_{1j} + \beta_1) \log(x_i) + \sum_{k=2}^7 \beta_k S_{ik} + \beta_8 B_i + \varepsilon_{ij} \\
 \begin{pmatrix} \alpha_{0j} \\ \alpha_{1j} \end{pmatrix} &\sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\alpha 0}^2 & \rho \sigma_{\alpha 0} \sigma_{\alpha 1} \\ \rho \sigma_{\alpha 0} \sigma_{\alpha 1} & \sigma_{\alpha 1}^2 \end{pmatrix} \right) \\
 \varepsilon_{ij} &\sim N(0, \sigma_{\varepsilon}^2),
 \end{aligned} \tag{1}$$

where β_1 was the coefficient associated with the log of the photo counts, β_k were coefficients associated with six species S_{ik} , $k=2,\dots,7$, and β_8 was the coefficient associated with Beaufort sea state B_i . Two random effects, α_{0j} and α_{1j} , allowed the relationship between $\log(y_{ij})$ and $\log(x_i)$ to vary among observers, α_{0j} in terms of the intercept and α_{1j} in terms of the slope coefficient β_1 . The two sets of random-effects coefficients had means of zero, variances $\sigma_{\alpha_0}^2$ and $\sigma_{\alpha_1}^2$, and correlation ρ . The assumption was that the 59 observers were a random selection from a larger pool of possible observers whose group size estimation tendencies were normally distributed.

Species S_{ik} entered the model as an indicator variable, with a value of 1 if group i was species k and 0 otherwise. Species were recorded in the field at the lowest possible taxonomic level, including subspecies. We combined the field identifications into six species categories: pantropical spotted dolphins (*Stenella attenuata*, 51 groups), spinner dolphins (*S. longirostris*, 40 groups), mixed spotted-spinner dolphin groups (78 groups), striped dolphins (*S. coeruleoalba*, 114 groups), common dolphins (*Delphinus delphis* and *D. capensis*, 87 groups) and other (64 groups). “Other” was a heterogeneous category including Risso’s dolphins (*Grampus griseus*), common bottlenose dolphins (*Tursiops truncatus*), rough-toothed dolphins (*Steno bredanensis*), short-finned pilot whales (*Globicephalus macrorhynchus*), and other groups which did not fit into the previous categories, such as mixed common-striped dolphin groups. In the eastern tropical Pacific Ocean, mixed spotted-spinner dolphin groups are common, so we included these as a distinct category. Sea state B was recorded on the Beaufort scale as an integer from 0 to 5; however, only one of the 434 calibration schools occurred in Beaufort 5 conditions, so the effective range of the model was Beaufort 0-4. Because the Beaufort scale is ordered, we modeled sea state as a continuous variable with a single linear coefficient. Models with sea state

as a categorical variable are addressed in the Discussion. S_i and B_i were the same for all observers for a given group i , so we omitted subscript j for these covariates in Eq. 1.

We considered four variants of Eq. 1 as candidate models: model 1, without species or sea-state effects ($\beta_k=0$ and $\beta_8=0$); model 2, with species but without sea-state effects ($\beta_8=0$); model 3, with sea-state but without species effects ($\beta_k=0$); and model 4, the full model with both species and sea-state effects. All four models included observers as a random effect.

Bayesian inference

To include model selection in a Bayesian framework, we fitted the models in R using reversible jump Markov Chain Monte Carlo (RJMCMC) methods (King *et al.* 2009, Oedekoven *et al.* 2014). In this approach, the model itself was treated as an additional parameter to be estimated, and the joint posterior distribution included both parameters and models (Appendix 1). A uniform discrete prior was specified for the four models, and uniform continuous priors were specified for all coefficients β and standard deviations σ in Eq. 1. Model probabilities were calculated as the fraction of iterations of the RJMCMC chain in each model after burn-in (Appendix 1).

The four models were also fitted in the BUGS language (Lunn *et al.* 2000) and compared with the Watanabe-Akaike (or Widely Applicable) Information Criterion (WAIC) (Watanabe 2010). WAIC can be viewed as an improvement to the Deviance Information Criterion (DIC) (Spiegelhalter *et al.* 2002), which has some shortcomings for hierarchical models (Plummer 2008, Millar 2009, Lunn *et al.* 2013). WAIC was calculated using pointwise predictive density at the observer level from the MCMC posterior samples for each model (Gelman *et al.* 2014, Vehtari *et al.* 2016). We used standard procedures to assess burn-in, autocorrelation, and

convergence of the MCMC samples (Appendix 2). BUGS code is given in Appendix 3 and R code for the RJMCMC analysis in Appendix 4.

For prediction, we sampled from the BUGS posterior samples of model 2, which had the most support (see Results). We made two kinds of predictions: an observer estimate given true group size and true group size given an observer estimate. For each, we predicted conditionally and unconditionally on both observer and species. A conditional prediction for an observer or species meant a prediction given that it was made by a particular observer or given that it was made of a group of a particular species. An unconditional prediction was calculated to infer estimation tendencies for a new (out-of-sample) observer and/or species – that is, estimation tendencies integrated over observers or species effects. Unconditional predictions were approximated by sampling observers and/or species randomly. We sampled the MCMC chain 50,000 times with replacement, each time also randomly sampling an observer and a species for unconditional inference. For the model error term σ_{ϵ}^2 , we made random draws from normal distributions with the MCMC sample variances. To preserve the covariance structure, we used the whole set of parameter values for each selected MCMC iteration, and computed observer estimate y given group size, or group size x given observer estimate, based on Eq. 1. We checked the accuracy of our predictions by comparing them to the photo counts (Appendix 5). For each of the 2,435 observer estimates, we determined if the central 95% credibility interval of predicted size included the photo count.

Results

Calibration schools

Group sizes of calibration schools ranged from 5 to 6,012 (Fig. 1). The set of 434 calibration schools represented about 8% of dolphin groups of the same species detected during the 11 surveys. On average, calibration schools were larger in size (because we did not photograph groups containing only a few dolphins) and were photographed in lower Beaufort sea states (because it was harder to obtain clear images in windy conditions) than for all dolphin groups. Importantly, the variation among independent observer estimates for a dolphin group was similar for calibration schools (mean CV 0.42, interquartile range 0.29-0.50) and all detected groups (mean CV 0.39, interquartile range 0.24-0.51). The number of calibration schools per observer ranged from 6 to 159, with a median of 33 and a mean of 41.3.

Observer estimates of dolphin group size

The raw data indicated that observers generally tended to underestimate dolphin group size; 69% of observer estimates were less than the photo count (Fig. 1). Both model selection methods indicated that the accuracy of observer estimates was affected by the species of the group but less so by Beaufort sea state. Posterior model probabilities indicated by the RJMCMC chain were 0.0, 0.984, 0.0, and 0.016 for models 1-4, respectively (Fig. 2). With proper selection of proposal distributions, models 2 and 4 had stationary distributions throughout the history of the chain (Fig. A1 in Appendix 1). WAIC scores showed a similar pattern favoring model 2 but with some support for model 4, with values of 3766.0, 3596.0, 3769.3, and 3598.5 for models 1-4, respectively.

Marginal posterior distributions of parameters for models 2 and 4 were similar (Table 1). For model 4, the sea-state coefficient β_8 was small in absolute value and the 95% credibility interval included 0, further indications that wind conditions in the range Beaufort 0-4 had little

effect on the accuracy of group size estimation. The coefficient for $\log(\text{photo count})$, β_1 , was <1.0 (mean 0.80, 95% credibility interval 0.76 to 0.83 for model 2), which meant that the tendency to underestimate increased with group size. Species coefficients decreased in the order mixed spotted-spinner, common, spotted, spinner, striped, and other (Fig. 3A). However, the posterior distributions of species coefficients overlapped (Table 1), indicating that the differences among species were modest. The random-effects coefficients were negatively correlated (mean $\rho = -0.79$, 95% credibility interval -0.59 to -0.91).

Observers differed in accuracy of group size estimation (Fig. 3B). Among the 59 observers, some tended to underestimate and others tended to overestimate. For spotted dolphin groups of 25, 50, 100, and 500 animals, the observers with the lowest estimation tendency had mean posterior estimates of 18, 29, 45 and 132, respectively, while the observers with the highest estimation tendency had mean posteriors of 42, 72, 125, and 585 (Table 2). The “average observer” (actually four different observers, one for each of the four group sizes in Table 2) had estimates of 25, 44, 78, and 290 for spotted dolphin groups of 25, 50, 100, and 500, respectively. Thus, over all observers, groups of 25 spotted dolphins were estimated accurately on average, but the range among observers was from underestimation by 29% to overestimation by 66% (Table 2). There were similarly large ranges in accuracy among observers for larger groups: for groups of 50, -43% to +45%; for groups of 100, -55% to +25%; and for groups of 500, -74% to +17%. The “average observer” underestimated spotted dolphin groups of 50, 100, and 500 animals by 11%, 22%, and 42%, respectively. We chose spotted dolphins for these numerical comparisons because spotted dolphins were near the middle of the species effect (Table 1). There would be less underestimation of group size for common dolphins and mixed groups of spotted-spinner dolphins, and more underestimation of group size for spinner, striped, and other

dolphins. Averaged over all species, the mean figures of underestimation were <1%, 16%, 27%, and 47% for groups of 25, 50, 100, and 500 animals.

The random-effects model allowed intercept and slope parameters to be estimated for each observer (Fig. 4), constrained by the hierarchical assumptions of normal distributions and correlation between slope and intercept. Visually, the greater importance of the observer effect relative to the species effect can be judged by comparing Fig. 3B with Fig. 3A. Numerically, the range of plausible values for observer intercepts ($1.5 \approx \pm 2 \sigma_{\alpha 0}$) was greater than the range of species effects (≈ 0.4), based on the mean posterior values in Table 1.

Accuracy decreased with group size (Fig. 5). Groups of 25 spotted dolphins were slightly underestimated, but groups of 500 were severely underestimated. For groups of 25, 50, 100, and 500 dolphins, posterior means for an out-of-sample observer (gray lines in Fig. 5) were 24.2, 42.2, 73.1, and 264.4, respectively. To show conditional estimates, we used observer #53 as an example. The black lines in Fig. 5 for observer #53 were slightly to the left of the gray lines unconditioned on observer, indicating that this observer tended to underestimate more than the average over all observers.

The posterior distributions of observer estimates were approximately normal on a natural logarithmic scale (Fig. 5). The distributions were quite wide, illustrating the high uncertainty (or low precision) in observer estimates of group size. Conditional estimates had higher precision than unconditional estimates. Estimates made by a particular observer (observer #53) for a particular species (spotted dolphins) had slightly higher precision (less uncertainty) than estimates by the same observer for an unknown species (compare thin dashed with thick solid black lines in Fig. 5). Unconditional estimates for any observer or species had the least precision

(thick gray lines in Fig. 5). The differences between conditional and unconditional estimates were small, however, in the context of the overall high variability of group size estimates.

Predictions of dolphin group size from observer estimates

Conversely, given an observer estimate, predicted true group size was usually larger than the estimate, especially for larger groups (Fig. 6). For observer estimates of 25, 50, 100, and 500 dolphins, posterior means were 26.0, 63.5, 154.0, and 1,194.5, respectively, for an out-of-sample observer (gray lines in Fig. 6). As with posterior distributions of observer estimates given group size, predicted group sizes conditional on observer and species had higher precision than unconditional estimates (compare black and gray lines in Fig. 6). Because observer #53 tended to underestimate more than average, predicted group size was larger for this observer than for the average over all observers.

Dolphin group size predicted from an observer estimate had high uncertainty. Coefficients of variation for predicted group size conditional on species ranged from approximately 0.7 to 0.9 (Table 3). Coefficients of variation for unconditional predictions were even larger, ranging from 0.9 to 1.9, due to the additional uncertainty of predicting group size for an unknown species. Given an observer estimate of 100 dolphins, the 95% credibility interval for the true size of the group ranged from 43 to 621 for a group of spotted dolphins, and from 37 to 776 for a group unconditional on species. Posterior distributions accurately captured the uncertainty in predicting dolphin group size from an observer estimate (Appendix 5, Fig. A4).

The degree to which an observer estimate was increased to estimate true group size depended on species. For an observer estimate of 25 dolphins, for example, the median predicted group size was smaller than 25 for mixed spotted-spinner and common groups, and

larger than 25 for spotted, spinner, striped, and other groups (Table 3). Because the exponentiated posterior distributions were lognormal, means were larger than medians. Therefore, we used the median (50% quantile) as the best measure of central tendency for these distributions, because there was equal probability of a value being higher or lower than the median. Integrated over species and observer effects, estimates of 25, 50, 100, and 500 were increased by 4%, 24%, 47%, and 122%, respectively, to obtain the medians of the posterior distributions of predicted group size (Table 3). In other words, given an observer estimate of 500 dolphins, the most probable true size of the group would be more than twice that number.

Discussion

Accuracy and precision

The discrepancy between an observer estimate of dolphin group size and the true number can be discussed in terms of two components: accuracy and precision. Accuracy is measured by the difference between the true number and the mean of repeated observations. Inaccurate measurement of group size leads to biased results. Precision is assessed by the random error among repeated observations. Random error will be positive for some observations and negative for others, but with a mean of zero. Low precision means high variance and greater uncertainty in results.

We found that accuracy of dolphin group size estimates depended on group size, observer, and species. Within the Beaufort 0-4 range of the calibration schools, Beaufort sea state had less effect on accuracy, once group size and observer effects had been accounted for.

There was a general tendency to underestimate dolphin group size, and this tendency increased with group size. The coefficient of the log of photo count (β_1 , Table 1) was < 1.0 , which meant that large groups were underestimated more than small groups. Observer estimates were accurate (on average) for dolphin groups of 25 animals, but were too low by 16% for groups of 50, too low by 27% for groups of 100, and too low by 47% for groups of 500 (Fig. 5). These estimates of accuracy averaged over all observers do not measure the accuracy of a particular observer, nor the discrepancy between an observer estimate and true group size for a particular group. Accuracy of dolphin group size estimation in this study applies within the range of calibration school sizes with a reasonable number of samples, roughly between 10 and 1000 animals (Fig. 3).

These results were broadly consistent with previous studies which showed that humans tend to underestimate group sizes in wildlife studies (Caughley 1974, Bayliss and Yeomans 1990, Cogan and Diefenbach 1998). The rate of decline in accuracy with group size ($\beta_1 = 0.80$, Table 1) falls in the range of perceptual experiments measuring underestimation of the number of dots on paper (Krueger 1972). Underestimation of large groups may have a physiological basis related to eye movement; estimation of small groups (about 10 or fewer objects) does not have this negative bias and seems to involve a different perceptual mechanism (Binda *et al.* 2011).

The degree of underestimation also varied by species. For the six species categories in this study, dolphin group size estimates were lower in the order: mixed spotted-spinner, common, spotted, spinner, striped, and other (Table 1, Fig. 3A). This order of species coefficients corresponded roughly to mean group size among the six species groups, with mixed spotted-spinner and common dolphin groups being largest, and striped and other dolphin groups

smallest. This correspondence suggests that the effects of group size and species were somewhat confounded.

Accuracy varied among the 59 observers. While there was an overall tendency to underestimate dolphin group size, some observers had a stronger tendency to underestimate, while others had a tendency to overestimate (Table 2). The random-effects model allowed the estimation of separate effects for each observer (Fig. 3B), but connected the observers as a group and allowed the tendency of all observers together to support estimation for each single observer (Fig. 4). A random-effects model is often understood in terms of “partial pooling.” It represents an intermediate approach between complete pooling (treating all observers as a single group, Fig. 1A) and no pooling (treating each observer independently, Fig. 1B). The random-effects approach spans a range of models between these extremes, and includes complete pooling and complete separation as special cases at the limits (Gelman and Hill 2007). The degree of pooling is related to the amount of shrinkage of individual effects toward the mean (Gelman and Pardoe 2006).

Precision of observer estimates of dolphin group size was strikingly low (Fig. 5). For a group of 100 dolphins, for example, estimates could range from about 30 to 200 with 95% probability. Regardless of an observer’s accuracy, it was common for the observer to estimate 50% high for one group and 50% low for the next. As a consequence, there was high variability among the independent observer estimates of group size, both for calibration schools as well as for non-calibration dolphin groups. The mean CV among observer estimates was 0.4 across a wide range of group sizes. Clearly, estimating the size of a dolphin group is a challenging task.

Statistical issues

As a measure of true group size, we used the mean of photo counts by three independent readers. A binomial moment estimator has been proposed for repeated counts with imperfect detection, *i.e.*, false negatives (DasGupta and Rubin 2005, Walsh *et al.* 2009), but in our study variation among counts of the three readers was also due to false positives. Large tuna, which frequently accompany dolphin groups in the eastern tropical Pacific, can be mistaken for a submerged dolphin in the photographs. Splashes and reflections might also be counted as a partially hidden dolphin.

RJMCMC and WAIC are two fully Bayesian approaches to model selection (Hooten and Hobbs 2015). RJMCMC treats the model itself as an additional unknown parameter to be estimated, while WAIC is a score function based on the predictive ability of the model. Both indicated that the accuracy of dolphin group size estimates varied by observer and species (model 2). There was little posterior support for model 4, which included Beaufort sea state (Fig. 2). The posterior odds of models 2 and 4 (the Bayes factor, Kass and Raftery 1995) was 60.6, indicating strong support for model 2 over model 4. The WAIC difference of 2.5 also indicated support of model 2 over model 4. If sea state was modeled as a categorical variable, model 4 had a posterior probability of zero (it was never selected in the RJMCMC algorithm), but if sea state was modeled as a continuous variable, model 4 was selected 2% of the time (Fig. 2, Fig. A1 in Appendix 1). Thus, it appeared that modeling sea state as a continuous variable rather than as separate factor variables was a more parsimonious approach. As there was little support for model 4, and because parameter estimates were similar for models 2 and 4 (Table 1), we focused on model 2 for inference and did not use model-averaged estimates.

Because Bayesian inference is based on conditional probabilities, it was possible to make inference regardless of observer and/or species, by integrating over observer and species effects.

The estimation tendency of a new, out-of-sample observer included the uncertainty of not knowing which observer, out of the “universe” of possible observers with different estimation tendencies, might be chosen. Such estimates unconditional for observer and species are shown as gray lines in Figs. 5 and 6. The greater uncertainty of the unconditional estimates is indicated by the wider probability distributions in those figures, relative to the conditional estimates shown with black lines.

Application of results

To obtain the best estimates of group size, we can use the estimation tendencies revealed in this study to adjust observer estimates of dolphin group size. We wish to predict true group size, given an observer estimate. The Bayesian approach allowed us to solve this inverse problem with proper accounting of uncertainty. Since, for groups larger than about 25 dolphins, there was a tendency to underestimate group size, predictions of true group size tended to be larger than the estimate (Fig. 6). Because the degree of underestimation depended on group size, species, and observer, the amount that a group size estimate had to be increased to predict true group size also depended on group size, species, and observer (Table 3). The amount that an estimate had to be increased could be substantial. For example, a group size estimate of 100 dolphins had to be increased by 47% to obtain the unconditional best (median) estimate of true group size.

Because an estimate of group size had low precision, predicted group size based on an estimate also had low precision. Posterior distributions had CVs of approximately 0.7 to 0.9 for groups of known species, and 0.9 to 1.9 for groups of any species (Table 3). For an out-of-sample observer estimate of 25 dolphins, for example, median predicted true group size was 25.9

animals (accuracy was good), but the 50% credibility interval extended from 16 to 42 dolphins, and the 95% credibility interval from 6 to 111 dolphins (Table 3). This source of uncertainty is usually ignored in distance sampling analyses, although Gerrodette and Forcada (2005) included uncertainty in group size through a bootstrap procedure. Most line-transect analyses compute the variance in expected group size from the sizes of the observed groups.

On cetacean line-transect surveys conducted by the Southwest Fisheries Science Center, three independent estimates of group size are recorded for each sighting. For the best estimate of group size, Gerrodette and Forcada (2005) used an average of the three calibration-adjusted observer estimates, weighted by the inverse of the group size estimation variance of each observer. The value of making several independent estimates of group size will be examined in a future paper.

Given our findings of inaccuracy for groups larger than 25 dolphins and low precision for groups of all sizes, it is worth noting that the estimates of group size in this study were a selected set of estimates made in optimal circumstances. Each group was approached with the specific objective of obtaining group size estimates, the observers had good views of the entire group, and the ship remained with the group until the observers had made their best possible estimates. Almost certainly the behavior of dolphin groups affects the accuracy and precision of group size estimates, but our set of calibration schools consisted of well-behaved groups that could be observed and photographed in their entirety.

Accuracy and precision may be lower for groups estimated in less optimal conditions. Schwarz *et al.* (2010) found that estimates of delphinid group sizes were 58% lower when the ship did not approach groups (passing mode) than when it did (closing mode). Barlow *et al.*

(1997) also found that group size estimates were smaller in passing mode. Barlow and Taylor (2005) found that an extended 90-min period of observation improved group size estimates of asynchronously diving sperm whales (*Physeter microcephalus*). The position of the observer may also matter. The estimates of group size in this study were made from a platform approximately 10 m above the water. The estimation tendencies reported here may not apply to other situations, such as estimates made from higher or lower platforms on a ship, or estimates made from land at various elevations and distances to sightings. Caughley *et al.* (1976) found that the accuracy of aerial counts varied with aircraft speed, height, and observer.

We conclude with two recommendations for studies that depend on estimates of cetacean group size. First, we recommend training to improve group size estimation. Although we were not able to measure how our pre-cruise training affected observers' estimates, we believe that the training had a positive effect. Training may include displays of groups of objects of known size, and instructions on estimating group size by counting subgroups of multiple animals. Second, we recommend assessment of accuracy and precision of group size estimation under the particular conditions of a study. The large budget of this study is unlikely to be replicated, but digital photography by drones is a more economical and much safer option today. Laake *et al.* (2012) used two observer teams to assess the accuracy of pod size estimates for migrating gray whales (*Eschrichtius robustus*). Although pod size was usually only one or two animals, correcting pod size estimates had an important effect on abundance estimates and inferred population trajectory.

If a study is unable to assess accuracy of group size estimates, the results of this study can be applied with appropriate caution. We have noted that biases might be different for group size estimates made under other conditions, such as greater distances. One of our central results was

449 that people varied widely in their group size estimation tendencies; therefore, the ideal is to
450 calibrate particular individual observers. However, the random-effects model for the observer
451 effect allowed inference for observers outside this study. Table 3 and Figures 5 and 6 show
452 posterior distributions for a new, out-of-sample observer – that is, accuracy and precision of
453 group size estimates which include the uncertainty of not knowing which observer, out of the
454 large number of possible observers with different estimation tendencies, might have been chosen.
455 Unless more specific information can be obtained, it would be reasonable to assume that the
456 estimation tendencies of the 59 observers in this study are representative of all observers.

457

Acknowledgments

458

459

460

461

462

463

We thank all the marine mammal observers who made the group size estimates, the aerial photographers and NOAA helicopter pilots, the captains and crews of the NOAA research vessels, the support staff at the Southwest Fisheries Science Center, and the Government of Mexico for flight permission in 2006. The photo counts were made by Jim Gilpatrick, Robin Westlake, Morgan Lynn, and Katie Cramer of the SWFSC. The manuscript benefitted from comments by Jim Gilpatrick, Jeffrey Moore, and two anonymous reviewers.

464

Literature cited

- 465 Barlow, J. 1997. Preliminary estimates of cetacean abundance off California, Oregon and
466 Washington based on a 1996 ship survey and comparisons of passing and closing modes.
467 Southwest Fisheries Science Center Administrative Report LJ-97-11. 25 p.
- 468 Barlow, J. and K. A. Forney. 2007. Abundance and population density of cetaceans in the
469 California Current ecosystem. *Fishery Bulletin* 105: 509-526.
- 470 Barlow, J. and B. L. Taylor. 2005. Estimates of sperm whale abundance in the northeastern
471 temperate Pacific from a combined acoustic and visual survey. *Marine Mammal Science*
472 21: 429-445.
- 473 Bayliss, P. and K. M. Yeomans. 1990. Use of low-level aerial photography to correct bias in
474 aerial survey estimates of magpie goose, and whistling duck density in the Northern
475 Territory. *Australian Wildlife Research* 17: 1-10.
- 476 Binda, P., M. Concetta Morrone, J. Ross and D. C. Burr. 2011. Underestimation of perceived
477 number at the time of saccades. *Vision Research* 51: 34-42.
- 478 Borchers, D. L. and K. P. Burnham. 2004. General formulation for distance sampling. Pages 6-
479 30 in S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers and L.
480 Thomas, eds. *Advanced distance sampling: Estimating abundance of biological*
481 *populations*. Oxford University Press, Oxford.
- 482 Caughley, G. 1974. Bias in aerial survey. *Journal of Wildlife Management* 38: 921-933.
- 483 Caughley, G., R. Sinclair and D. Scott-Kemmis. 1976. Experiments in aerial survey. *Journal of*
484 *Wildlife Management* 40: 290-300.

485 Cogan, R. D. and D. R. Diefenbach. 1998. Effect of undercounting and model selection on a
486 sightability-adjustment estimator for elk. The Journal of Wildlife Management 62: 269-
487 279.

488 DasGupta, A. and H. Rubin. 2005. Estimation of binomial parameters when both n, p are
489 unknown. Journal of Statistical Planning and Inference 130: 391-404.

490 Erwin, R. M. 1982. Observer variability in estimating numbers: an experiment. Journal of Field
491 Ornithology 53: 159-167.

492 Gelman, A. and J. Hill 2007. Data analysis using regression and multilevel/hierarchical models.
493 Cambridge University Press, Cambridge, MA.

494 Gelman, A., J. Hwang and A. Vehtari. 2014. Understanding predictive information criteria for
495 Bayesian models. Statistics and Computing 24: 997-1016.

496 Gelman, A. and I. Pardoe. 2006. Bayesian measures of explained variance and pooling in
497 multilevel (hierarchical) models. Technometrics 48: 241-251.

498 Gelman, A., G. O. Roberts and W. R. Gilks. 1996. Efficient Metropolis jumping rules. Bayesian
499 Statistics 5: 599-607.

500 Gerrodette, T. and J. Forcada. 2005. Non-recovery of two spotted and spinner dolphin
501 populations in the eastern tropical Pacific Ocean. Marine Ecology Progress Series 291: 1-
502 21.

503 Gilpatrick, J. W., Jr. 1993. Method and precision in estimation of dolphin school size with
504 vertical aerial photography. Fishery Bulletin 91: 641-648.

505 Green, P. J. 1995. Reversible jump Markov Chain Monte Carlo computation and Bayesian model
506 determination. Biometrika 82: 711-732.

507 Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their
 508 applications. *Biometrika* 57: 97-109.

509 Hooten, M. B. and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists.
 510 *Ecological Monographs* 85: 3-28.

511 Kass, R. E. and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical*
 512 *Association* 90: 773-795.

513 King, R., B. J. T. Morgan, O. Gimenez and S. P. Brooks 2009. Bayesian analysis for population
 514 ecology. Chapman & Hall/CRC.

515 Krueger, L. E. 1972. Perceived numerosity. *Perception & Psychophysics* 11: 5-9.

516 Laake, J. L., A. E. Punt, R. Hobbs, M. Ferguson, D. Rugh and J. Breiwick. 2012. Gray whale
 517 southbound migration surveys 1967-2006: An integrated re-analysis. *Journal of Cetacean*
 518 *Research and Management* 12: 287-306.

519 Lunn, D., C. Jackson, N. Best, A. Thomas and D. Spiegelhalter 2013. The BUGS book: A
 520 practical introduction to Bayesian analysis. CRC Press, Boca Raton, FL.

521 Lunn, D. J., A. Thomas, N. Best and D. Spiegelhalter. 2000. WinBUGS -- a Bayesian modelling
 522 framework: Concepts, structure, and extensibility. *Statistics and Computing* 10: 325-337.

523 Metropolis, N., A. Rosenbluth, M. N. Rosenbluth and A. Teller. 1953. Equation of state
 524 calculations by fast computing machines. *The Journal of Chemical Physics* 21: 1087-
 525 1092.

526 Millar, R. B. 2009. Comparison of hierarchical Bayesian models for overdispersed count data
 527 using DIC and Bayes' factors. *Biometrics* 65: 962-969.

- Oedekoven, C. S., S. T. Buckland, M. L. Mackenzie, R. King, K. O. Evans and L. W. Burger, Jr. 2014. Bayesian methods for hierarchical distance sampling models. *Journal of Agricultural, Biological, and Environmental Statistics* 19: 219-239.
- Plummer, M. 2008. Penalized loss functions for Bayesian model comparison. *Biostatistics* 9: 523-539.
- Ryan, P. G. and J. Cooper. 1989. Observer precision and bird conspicuousness during counts of birds at sea. *South African Journal of Marine Science* 8: 271-276.
- Schwarz, L. K., T. Gerrodette and F. I. Archer. 2010. Comparison of closing and passing mode from a line-transect survey of delphinids in the eastern Tropical Pacific Ocean. *Journal of Cetacean Research and Management* 11: 253-265.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin and A. Van Der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, B* 64: 583-639.
- Vehtari, A., A. Gelman and J. Gabry. 2016. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 26: 1-20.
- Walsh, D. P., C. F. Page, H. Campa, Iii, S. R. Winterstein and D. E. Beyer, Jr. 2009. Incorporating estimates of group size in sightability models for wildlife. *The Journal of Wildlife Management* 73: 136-143.
- Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11: 3571-3594.

550 Table 1. Marginal posterior distributions of parameters for the two models with posterior
551 support. Distributions are summarized by means, standard deviations (SD) and three quantiles.
552 All parameters had uniform prior distributions. RE = random effects for observers. See Eq. 1
553 for definitions of parameters.

Parameter	Model 2					Model 4				
	Mean	SD	2.5%	50%	97.5%	Mean	SD	2.5%	50%	97.5%
Log(photo count), β_1	0.796	0.018	0.761	0.796	0.832	0.797	0.018	0.760	0.798	0.831
Sp: spotted-spinner, β_2	0.805	0.087	0.635	0.803	0.974	0.818	0.087	0.653	0.816	0.995
Sp: common, β_3	0.757	0.084	0.592	0.755	0.926	0.776	0.087	0.613	0.774	0.951
Sp: spotted, β_4	0.656	0.082	0.497	0.654	0.816	0.674	0.083	0.521	0.672	0.843
Sp: spinner, β_5	0.603	0.088	0.433	0.604	0.778	0.620	0.091	0.448	0.618	0.806
Sp: striped, β_6	0.513	0.075	0.370	0.511	0.660	0.531	0.078	0.386	0.530	0.687
Sp: other, β_7	0.423	0.076	0.273	0.424	0.576	0.442	0.079	0.291	0.440	0.601
Beaufort sea state, β_8	na	na	na	na	na	-0.010	0.010	-0.030	-0.010	0.010
SD intercept RE, $\sigma_{\alpha 0}$	0.382	0.073	0.245	0.377	0.534	0.385	0.073	0.252	0.381	0.540
SD slope RE, $\sigma_{\alpha 1}$	0.104	0.017	0.073	0.102	0.138	0.104	0.017	0.073	0.103	0.139
Correlation RE, ρ	-0.792	0.085	-0.912	-0.808	-0.586	-0.792	0.088	-0.912	-0.809	-0.578
SD model, σ_{ϵ}	0.497	0.008	0.482	0.497	0.512	0.497	0.007	0.482	0.497	0.511

554

555

556 Table 2. Summary of estimation tendencies among observers. For each group size, the entries in
557 the table show the distribution of the means of the posteriors of the 59 observers for estimates of
558 a group of spotted dolphins. “Mean diff.”, “Min diff.” and “Max diff.” are the differences
559 between the mean (or minimum or maximum) of the mean observer estimates and true group
560 size, expressed as percentages of group size.

Group size	Distribution of means of observer estimates						Mean diff.	Min diff.	Max diff.
	Mean	Min	25%	50%	75%	Max			
25	25.4	17.7	21.2	25.9	28.7	41.5	2%	-29%	66%
50	44.4	28.5	35.5	45.0	51.1	72.4	-11%	-43%	45%
100	77.7	44.9	60.3	78.3	89.1	124.6	-22%	-55%	25%
500	289.7	131.7	201.8	293.0	355	585.3	-42%	-74%	17%

561

562 Table 3. Predicted dolphin group sizes given observer estimates of 25, 50, 100, and 500 animals
563 by a new (out-of-sample) observer, for six dolphin species and integrated over species (“any
564 species”). Posterior distributions have been exponentiated to show values on the scale of the
565 number of dolphins. Distributions of predicted group size are approximately lognormal, and are
566 summarized by means, standard deviations (SD), coefficients of variation (CV) and five
567 quantiles. “Difference” is the difference between the median (the 50% quantile) of predicted
568 group size and observer estimate, expressed as a percentage of the observer estimate.

569

Observer estimate	Dolphin species	Predicted group size								Difference
		Mean	SD	CV	2.5%	25%	50%	75%	97.5%	
25	spotted-spinner	26.4	19.6	0.74	5.3	13.3	21.2	33.3	78.2	-15%
	common	28.2	21.1	0.75	5.5	14.3	22.6	35.8	83.4	-10%
	spotted	32.3	24.4	0.76	6.6	16.3	25.9	40.8	95.7	4%
	spinner	34.6	26.3	0.76	6.9	17.4	27.7	43.6	103.6	11%
	striped	39.0	29.1	0.75	7.9	19.9	31.4	49.3	113.9	26%
	other	43.8	33.3	0.76	8.8	22.1	35.2	55.3	130.3	41%
	any species	34.1	30.0	0.88	6.2	16.1	25.9	42.0	110.6	4%
50	spotted-spinner	66.6	50.4	0.76	13.9	33.8	53.5	83.6	196.6	7%
	common	70.7	53.2	0.75	14.7	36.1	56.7	88.7	209.9	13%
	spotted	81.1	61.6	0.76	17.0	41.2	65.1	101.4	239.7	30%
	spinner	87.3	66.7	0.76	18.1	44.3	70.3	110.1	256.5	41%
	striped	98.6	75.2	0.76	20.2	49.9	78.6	123.9	294.5	57%
	other	110.6	84.1	0.76	23.1	56.1	88.8	138.8	325.8	78%
	any species	84.5	79.6	0.94	15.7	38.8	62.1	102.6	286.9	24%
100	spotted-spinner	167.2	128.0	0.77	34.7	84.6	133.2	208.8	502.4	33%
	common	179.5	137.4	0.77	37.7	90.9	143.3	224.6	538.7	43%
	spotted	206.2	160.0	0.78	43.4	103.9	163.4	257.5	620.8	63%
	spinner	222.2	175.1	0.79	46.7	111.4	175.4	276.6	680.3	75%
	striped	247.3	191.0	0.77	52.7	124.6	195.8	308.9	748.2	96%
	other	281.7	219.6	0.78	59.4	142.1	223.3	349.7	847.5	123%
	any species	211.2	241.4	1.14	37.4	90.2	147.4	248.5	776.3	47%
500	spotted-spinner	1466.7	1215.5	0.83	305.0	723.5	1137.9	1810.4	4557.1	128%
	common	1581.4	1335.3	0.84	326.1	766.7	1218.6	1956.9	4981.8	144%
	spotted	1801.7	1518.8	0.84	367.9	876.1	1386.9	2227.2	5629.0	177%
	spinner	1941.8	1651.7	0.85	400.7	938.3	1490.8	2399.8	6192.8	198%
	striped	2181.9	1972.4	0.90	447.1	1057.2	1674.9	2689.5	6898.4	235%
	other	2483.6	2145.0	0.86	498.4	1190.7	1892.8	3053.4	7876.5	279%
	any species	1944.2	3600.3	1.85	250.6	626.9	1108.2	2103.6	8468.8	122%

Figure captions

Fig. 1: Dolphin group size calibration data plotted on logarithmic scales. (A) Photo counts and observer estimates of group size for 434 calibration schools. The size of each group was estimated independently by multiple (usually 6) shipboard observers. The dashed line is a regression of $\log(\text{observer estimate})$ on $\log(\text{photo count})$, while the solid gray line is a 1:1 relationship. (B) Regressions of $\log(\text{observer estimate})$ on $\log(\text{photo count})$ for each of the observers.

Fig. 2. Prior and posterior probabilities of four models of dolphin group size estimation based on RJMCMC. Differences among observers were modeled as random effects (RE) in all four models; species and sea state were fixed effects.

Fig. 3. Estimates of (A) species and (B) observer effects on dolphin group size estimation. Regression lines are based on means of posterior distributions.

Fig. 4. Posterior distributions of random effects for each observer for (A) intercept α_0 and (B) slope α_1 (see Eq. 1). Points are means and lines are central 95% credibility intervals.

Fig. 5. Posterior distributions of observer estimates for dolphin groups of 25, 50, 100, and 500 animals. Thin dashed lines are the distributions of estimates for a given observer (#53) whose tendencies were estimated in this study, for a given species (spotted dolphins). Thick black lines are the distributions for the same observer for any species (integrated over species). Thick gray lines are the distributions for a new, out-of-sample observer with unknown tendencies, for any species (integrated over observers and species). The probability densities (vertical scale) of all distributions are scaled relative to the maximum value.

595 Fig. 6. Predicted dolphin group sizes given observer estimates of 25, 50, 100, and 500 animals.
596 Thin dashed lines are the distributions of group size for a given observer (#53) whose tendencies
597 were estimated in this study, for a given species (spotted dolphins). Thick black lines are the
598 distributions for the same observer for any species (integrated over species). Thick gray lines are
599 the distributions for a new, out-of-sample observer with unknown tendencies, for any species
600 (integrated over observers and species). The probability densities (vertical scale) of all
601 distributions are scaled relative to the maximum density value.

602

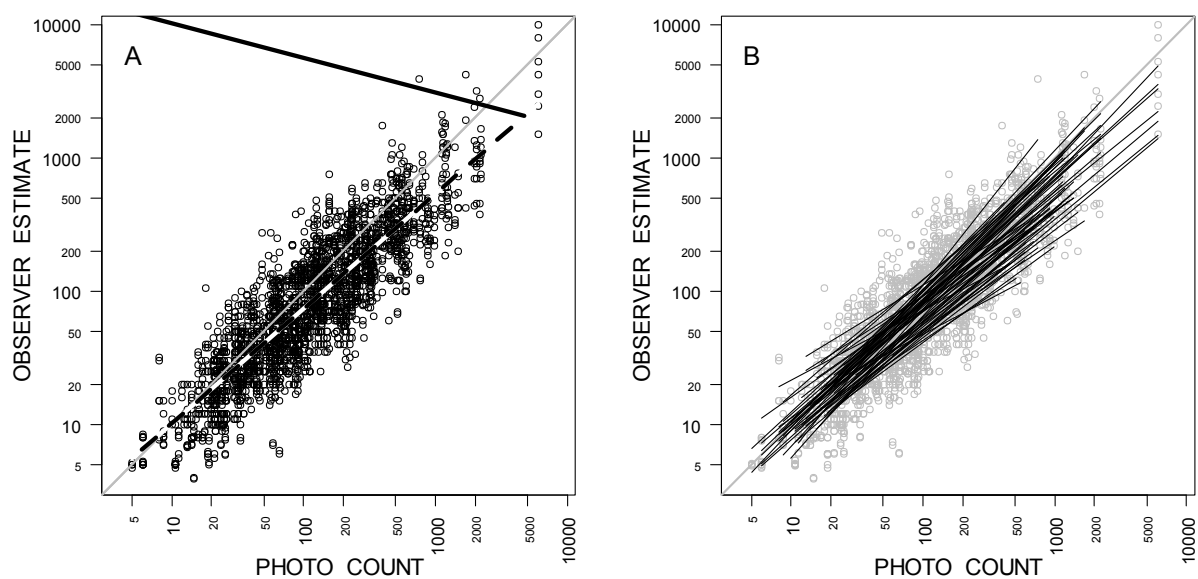


Fig. 1

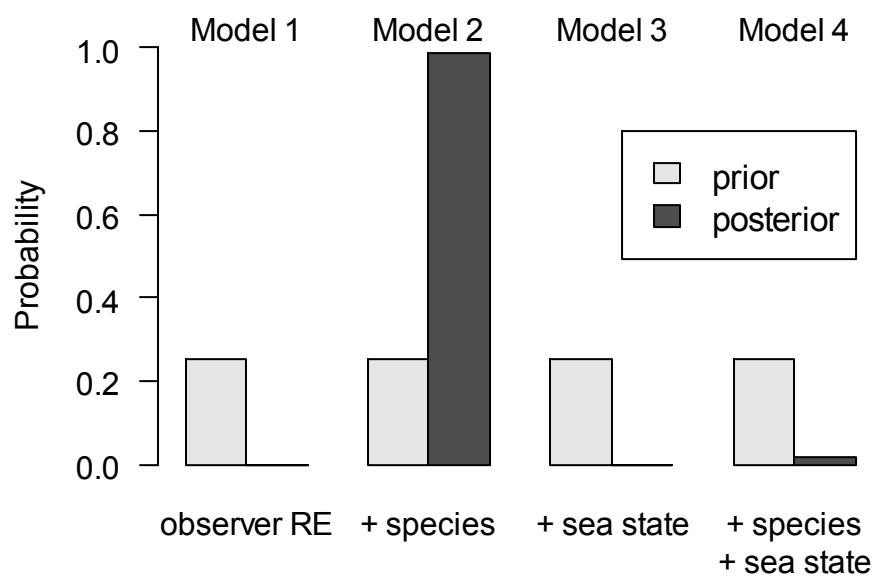


Fig. 2

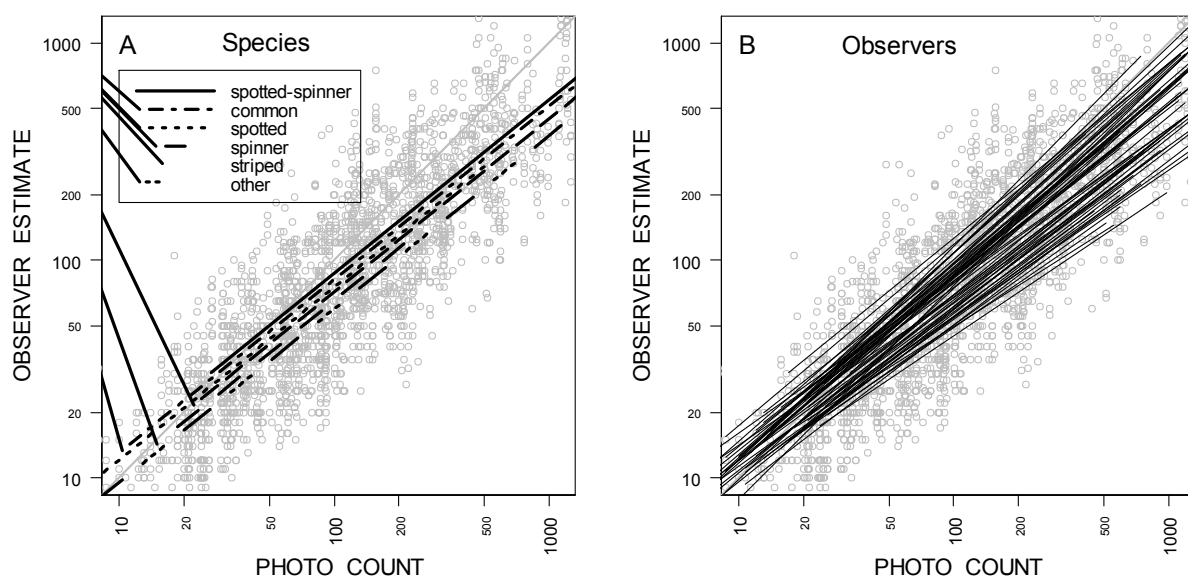


Fig. 3

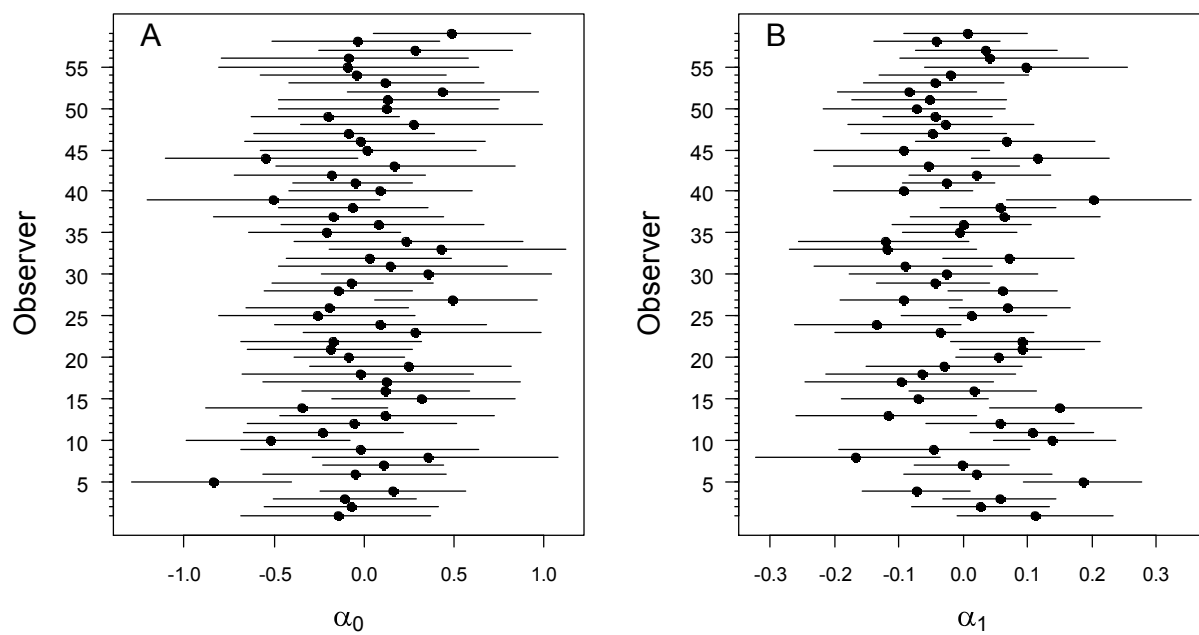


Fig. 4

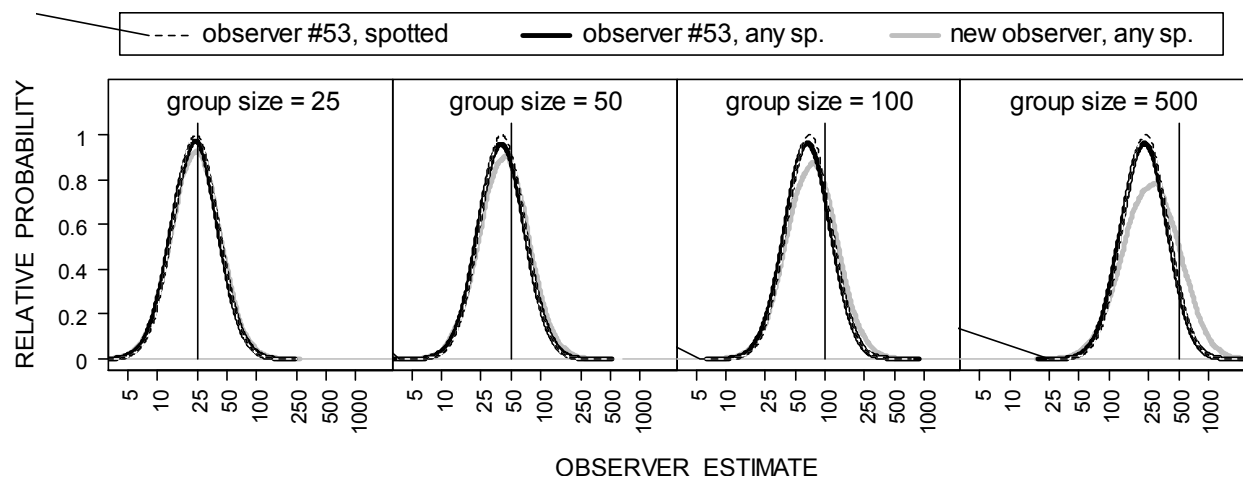


Fig. 5

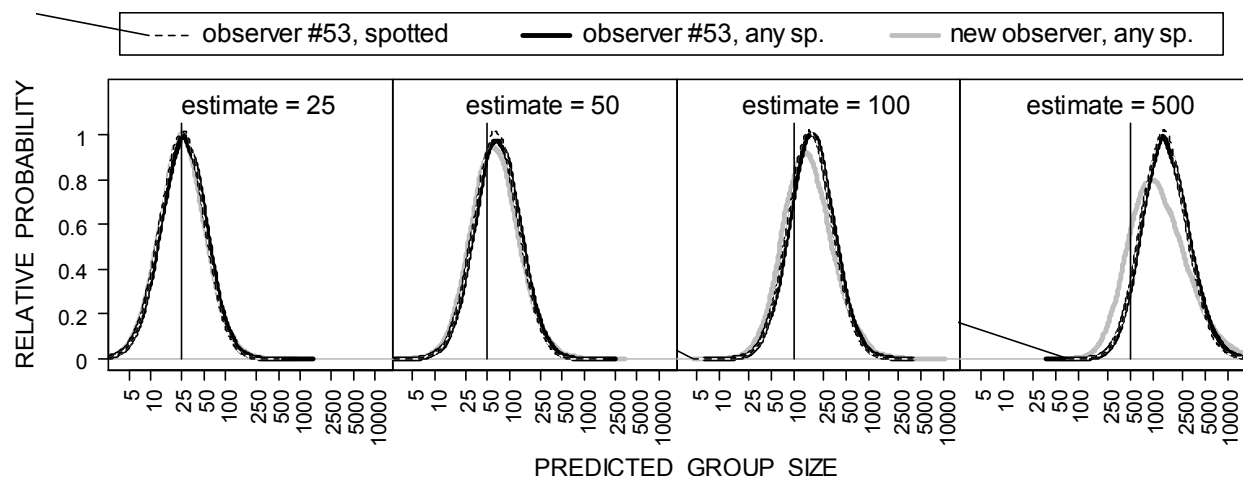


Fig. 6

Appendix 1. Reverse Jump Markov Chain Monte Carlo (RJMCMC)

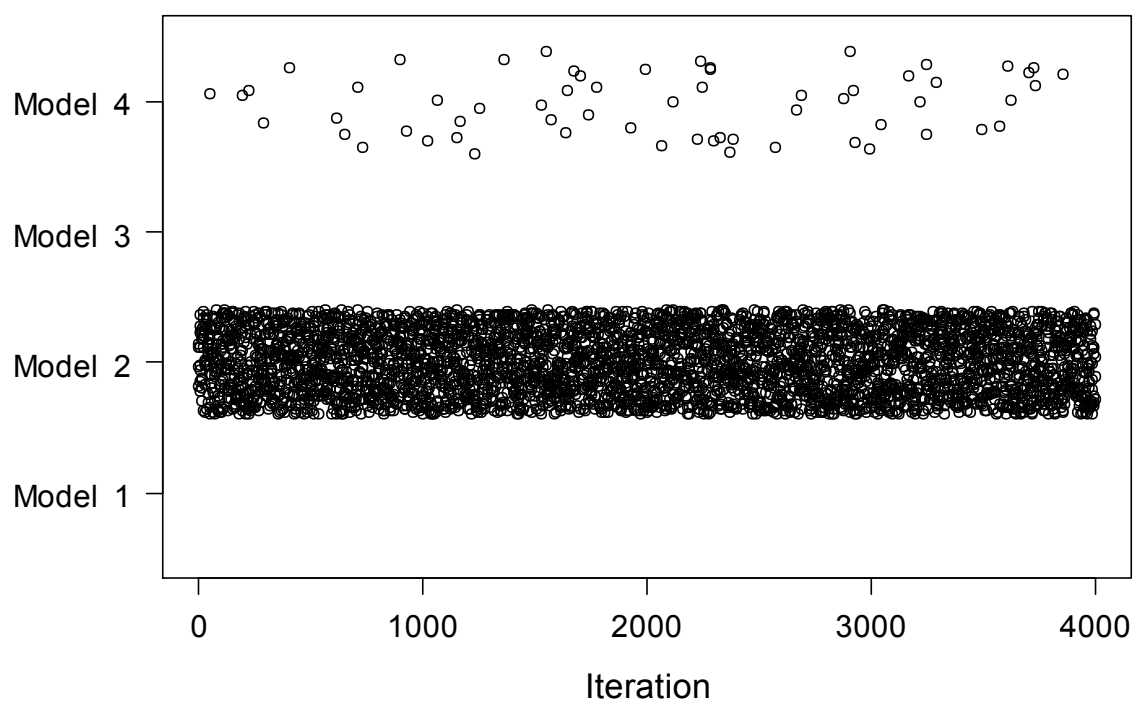
Random-effects models such as Eq. (1) can be implemented in a Bayesian framework using hierarchical models where each parameter, including the random-effects standard deviations, are assumed to have a distribution. Markov Chain Monte Carlo (MCMC) simulation can be used to obtain summary statistics of the posterior distributions of the parameters given the data. To include model selection in our analysis, we treated the model itself as a parameter and formed the joint posterior distribution of both parameters and models. An RJMCMC algorithm (Green 1995) explored this posterior distribution. The RJMCMC algorithm represented a random walk, where each iteration consisted of two steps: (1) the reversible jump (RJ) step where we proposed to move to a different model (the between-model move), and (2) the Metropolis-Hastings (MH) step where we updated the parameters from the current model (the within-model move). We placed uniform priors on all parameters with an upper bound of 1, and a lower bound of -1 for coefficients and a lower bound of 0 for standard deviations.

All models included in the analysis contained the intercept and the log of the photo counts as well as their corresponding random-effects coefficients (Eq. 1). Hence, the RJ step at each iteration consisted of proposing to add or delete each of the two remaining covariates (*sea-state* and *species*) in turn, depending on whether the covariate was in the current model or not. Four different models were possible that differed only in the inclusion or exclusion of species (β_k , with $k = 2, \dots, 7$) and sea state (β_8) coefficients in Eq.(1): for model 1, $\beta_k=0$ and $\beta_8=0$; for model 2, $\beta_8=0$; for model 3, $\beta_k=0$; and for model 4, both species and sea-state coefficients were non-zero (full model) . A proposal to add a covariate to the model involved drawing random samples from the respective proposal distributions for the parameters and accepting this proposal based on the calculated acceptance probability (see, *e.g.*, King *et al.* 2009 on how to obtain the

acceptance probability). A proposal to delete a covariate from a model involved setting its coefficients to zero and accepting this proposal based on the calculated acceptance probability. The four models were considered equally likely a priori.

The MH step at each iteration consisted of updating the parameters that were currently in the model using an MH update (Metropolis *et al.* 1953, Hastings 1970). This included the coefficient associated with the log of the photo counts, the standard deviations associated with the random effects and model errors as well as the coefficients for *species* and *sea-state* if these covariates were in the current model. Furthermore, all random-effects coefficients were updated during each iteration. In particular, this update involved a random walk single-update with normal proposal distributions, where the mean was equal to the current value of the parameter (or random-effects coefficient) and the standard deviations were fine-tuned during pilot tuning to achieve appropriate acceptance rates (Gelman *et al.* 1996).

The chain was started with the full model and completed 210,000 iterations. We discarded the first 10,000 as burn-in and thinned the chain by retaining every 50th value, thus obtaining a posterior sample of 4000 values. Posterior model probabilities were the fraction of iterations that the chain spent in the respective model. Models 1 and 3 were never selected; model 4 was selected 1.6% of the time consistently through the history of the chain (Fig. A1). Similar results were obtained regardless of which model was used to initiate the chain.

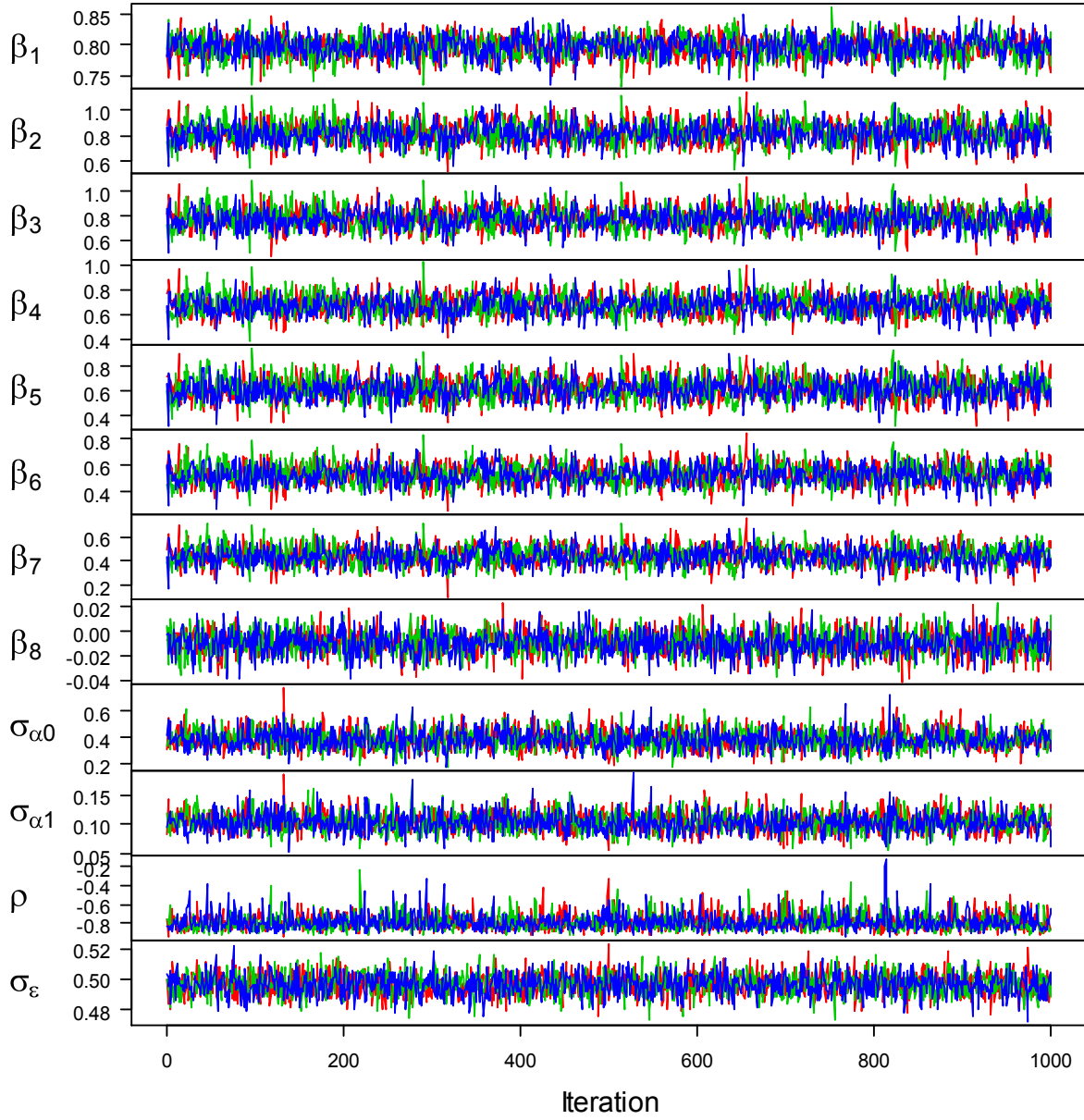


669
 670 Fig. A1. Sequence of RJMCMC jumps among models after burn-in. Results were similar
 671 regardless of which model was chosen to initiate the chain. To show separate points, random
 672 values have been added to each point (jittering).

673

Appendix 2. BUGS models and diagnostics for MCMC sampling

Each of the four variants of Eq. 1 was implemented in the BUGS language. Uniform priors were specified for all parameters except the random-effects coefficients, which were latent. Due to the large amount of data, specification of other priors, such as normal distributions (lognormal distributions for variance parameters) with means far from values supported by the data, had no effect on posterior distributions. For each model, we ran three chains of 120,000 iterations each, discarding the first 20,000 as burn-in from different random initial starting values. For the remaining 100,000 iterations, we retained every 100th value (thinning) to reduce autocorrelation. Thus the final sample consisted of 1000 values for each of three chains. The effective sample size for each parameter, calculated with R package coda, was near 1000, indicating that autocorrelation was low. The chains were well-mixed for all parameters (Fig. A2), and converged to similar values (Fig. A3).



689

690

691

692

Fig. A2. Traces of posterior samples. Green, red and blue lines show three independent MCMC chains of 1000 iterations each, with different initial values. See Eq. 1 and Table 1 for definitions of parameters.

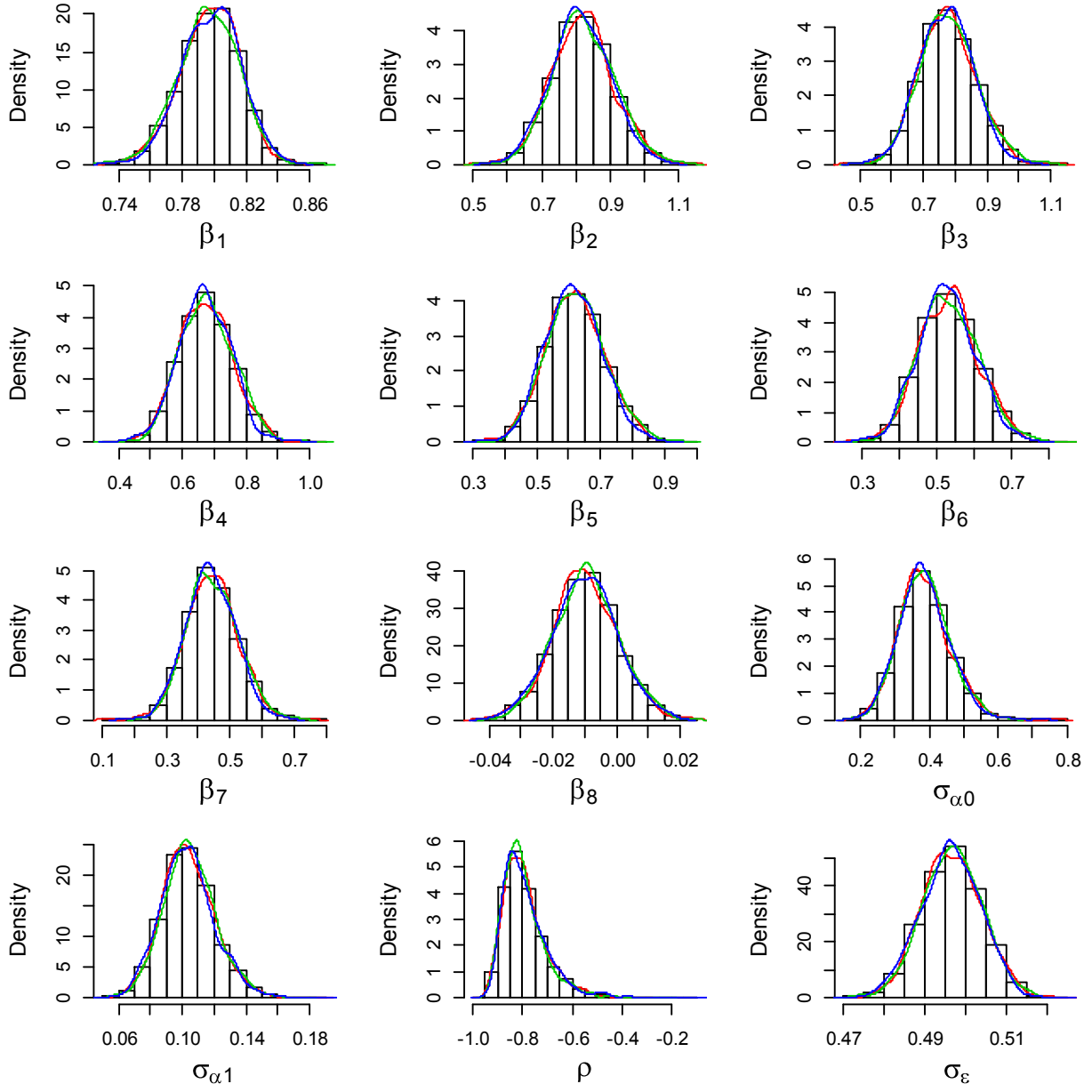


Fig. A3. Marginal posterior probability density distributions for parameters. Green, red and blue lines show three independent MCMC chains with different initial values. The histogram is the total sample of all three chains. See Eq. 1 and Table 1 for definitions of parameters.

```

699  Appendix 3. BUGS code

700  BUGS.model <- function() {
701    for (i in 1:n) {
702      y[i] ~ dnorm(y.hat[i],tau.model)
703      y.hat[i] <- a0[obs[i]] + a1[obs[i]]*x[i] + b.sp[sp[i]]          # model 2
704      # y.hat[i] <- a0[obs[i]] + a1[obs[i]]*x[i] + b.sp[sp[i]] + b.bf*bf[i]  # model 4
705    }
706    tau.model <- pow(sigma.model,-2)
707    sigma.model ~ dunif(sigma.min,sigma.max)  # prior
708    for (i in 1:6) {b.sp[i] ~ dunif(b.min,b.max)}  # 6 species factor levels
709    # b.bf ~ dunif(b.min,b.max)                # sea state
710    for (j in 1:n.obs) {
711      a0[j] <- A[j,1]
712      a1[j] <- A[j,2]
713      A[j,1:2] ~ dmnorm(A.hat[j,],Tau.A[,])
714      A.hat[j,1] <- 0    # mean of intercept random effects
715      A.hat[j,2] <- b1    # mean of slope random effects
716    }
717    b1 ~ dunif(b.min,b.max)    # prior
718    Tau.A[1:2,1:2] <- inverse(Sigma.A[,])
719    Sigma.A[1,1] <- pow(sigma.a0,2)
720    Sigma.A[2,2] <- pow(sigma.a1,2)
721    Sigma.A[1,2] <- rho*sigma.a0*sigma.a1
722    Sigma.A[2,1] <- Sigma.A[1,2]
723    sigma.a0 ~ dunif(sigma.min,sigma.max) # prior
724    sigma.a1 ~ dunif(sigma.min,sigma.max) # prior
725    rho ~ dunif(-1,1)                # prior
726  }
727

```

```

728 Appendix 4. R code for RJMCMC analysis
729 # RJMCMC calibration analysis for ETP dolphin school size estimation
730 #
731 library(tcltk2) # for progress bar
732 ## Proposal distributions for parameters for RJ step
733 rjprop.mean.sp <- rep(0,5)
734 rjprop.mean.bft <- 0
735 rjprop.sd.bft <- 0.1
736 rjprop.sd.sp <- rep(0.3,5)
737 ## Proposal distributions for parameters for MH step
738 mhprop.sd.int <- 0.035
739 mhprop.sd.ph <- 0.007
740 mhprop.sd.sp <- rep(0.04,5)
741 mhprop.sd.bft <- 0.005
742 mhprop.sd.sd.model <- 0.01
743 mhprop.sd.sd.obs.int = 0.01
744 mhprop.sd.sd.obs.ph = 0.01
745 mhprop.sd.params <- c(mhprop.sd.int, mhprop.sd.ph, mhprop.sd.sp, mhprop.sd.bft,
746 mhprop.sd.sd.model, mhprop.sd.sd.obs.int, mhprop.sd.sd.obs.ph)
747 names(mhprop.sd.params) <- c('sd.int','sd.ph',rep('sd.sp',5),'sd.bft',
748 'sd.sd.model','sd.sd.obs.int','sd.sd.obs.ph')
749
750 ##### model set-up #####
751 ## Starting values for the parameters
752 # fixed effects
753 int.0 <- 0.7 # intercept
754 ph.0 <- 0.8 # slope for photo
755 sp.0 <- rjprop.mean.sp # factor covariate with 6 levels (first level absorbed in the intercept)
756 bft.0 <- rjprop.mean.bft # beaufort coefficient
757 sd.model.0 <- 0.5 # standard deviation of model errors
758 # random effects for observers
759 sd.obs.int.0 <- 0.2 # intercept for regression
760 re.obs.int <- rnorm(n.obs,0,sd.obs.int.0)
761 names(re.obs.int) <- sort(unique(observers))
762 sd.obs.ph.0 <- 0.05
763 re.obs.ph <- rnorm(n.obs,0,sd.obs.ph.0)
764 names(re.obs.ph) <- sort(unique(observers))
765 params <- c(int.0,ph.0,sp.0,bft.0,sd.model.0,sd.obs.int.0,sd.obs.ph.0)
766 names(params) <-
767 c('int','ph',paste("sp",levels(species)[2:6],sep="."),'bft','sd.model','sd.obs.int','sd.obs.ph')
768 param.list <- matrix(0,4,8)
769 param.list[1,c(1,2)] <- 1
770 param.list[2,c(1:7)] <- 1

```

```

771 param.list[3,c(1,2,8)] <- 1
772 param.list[4,1:8] <- 1
773
774 # choose the model
775 cur.mod <- 1
776 # which parameters are switched on
777 cur.p <- param.list[cur.mod,]
778 params[1:8] <- params[1:8]*cur.p
779
780 ## Prior limits for parameters
781 prior.params.lo <- -1
782 prior.params.hi <- 1
783 prior.sd.lo <- 0
784 prior.sd.hi <- 1
785
786 # number of iterations, about 3000 per hour
787 n.iter <- 3000*70      # total number of iterations
788 n.thin <- 10          # thinning; number of posterior samples will be floor(n.iter/n.thin) + 1
789
790 # setting up matrices that will store the posterior samples
791 nr <- round(n.iter/n.thin,0)+1 # number of rows is thinned no. of updates + starting value
792 params.mat <- matrix(NA,nr,length(params))
793 colnames(params.mat) <- names(params)
794 params.mat[1,] <- params
795 re.obs.int.mat <- matrix(NA,nr,n.obs)
796 colnames(re.obs.int.mat) <- paste("obs",levels(observers),".int",sep="")
797 re.obs.int.mat[1,] <- re.obs.int
798 re.obs.ph.mat <- matrix(NA,nr,n.obs)
799 colnames(re.obs.ph.mat) <- paste("obs",levels(observers),".ph",sep="")
800 re.obs.ph.mat[1,] <- re.obs.ph
801
802 # vector for storing model choices
803 model <- array(NA,nr)
804 # the predictor
805 x <- l.photo
806 # the response
807 y <- l.best
808
809 ##### the likelihood equations
810 log.lik <- function(y = y, x = x, params = params, re.obs.int = re.obs.int, re.obs.ph = re.obs.ph){
811   sp.params<-params[3:7]    # these will be zero if beaufort is not included in the model
812   bft.params<-params[8]    # these will be zero if species is not included in the model
813   mu <- params['int'] + re.obs.int[observers] + (params['ph'] + re.obs.ph[observers]) * x +
814   c(0,sp.params)[match(species,levels(species))] + bft.params[1]*beaufort

```

```

815   log.lik <- sum(log(dnorm(y,mu,params['sd.model']))) +
816   sum(log(dnorm(re.obs.ph,0,params['sd.obs.ph']))) +
817   sum(log(dnorm(re.obs.int,0,params['sd.obs.int'])),na.rm=T)
818   log.lik
819 }
820
821 # test
822 log.lik(y = l.best, x = l.photo, params = params, re.obs.int = re.obs.int, re.obs.ph = re.obs.ph)
823
824 #####
825 # progress bar
826 pb <- tkProgressBar(title = "progress bar", min = 0,max = n.iter, width = 200)
827
828 # the RJMCMC algorithm
829 isave <- 1      # set the counter; first value is starting value
830 for (b in 2:n.iter){
831   newparams <- params
832
833   ##### the RJ step
834   if(cur.p[3]==0){ # if species is currently not in the model, propose to add it
835     newparams[3:7] <- rnorm(5,rjprop.mean.sp,rjprop.sd.sp) ##### changed from
836     1 to 5
837     new.lik <- log.lik(y = y, x = x, params = newparams, re.obs.int = re.obs.int, re.obs.ph =
838     re.obs.ph)
839     cur.lik <- log.lik(y = y, x = x, params = params, re.obs.int = re.obs.int, re.obs.ph = re.obs.ph)
840     num <- new.lik + sum(log(dunif(newparams[3:7],prior.params.lo,prior.params.hi))) # add
841     priors for new parameters
842     den <- cur.lik + sum(log(dnorm(newparams[3:7],rjprop.mean.sp,rjprop.sd.sp))) # add
843     proposal densities for new parameters
844     A<-min(1,exp(num-den))
845     V<-runif(1)
846     ifelse(V<=A,{params[3:7]<-newparams[3:7];cur.p[3:7]<-1},{newparams[3:7]<-params[3:7]})
847   }
848   else{ # if species is currently in the model, propose to delete it
849     newparams[3:7] <- 0
850     new.lik <- log.lik(y = y, x = x, params = newparams, re.obs.int = re.obs.int, re.obs.ph =
851     re.obs.ph)
852     cur.lik <- log.lik(y = y, x = x, params = params, re.obs.int = re.obs.int, re.obs.ph = re.obs.ph)
853     num <- new.lik + sum(log(dnorm(params[3:7],rjprop.mean.sp,rjprop.sd.sp))) # add proposal
854     densities for current parameters
855     den <- cur.lik + sum(log(dunif(params[3:7],prior.params.lo,prior.params.hi))) # add priors for
856     current parameters
857     A<-min(1,exp(num-den))
858     V<-runif(1)

```

```

859   ifelse(V<=A,{params[3:7]<-newparams[3:7];cur.p[3:7]<-0},{newparams[3:7]<-params[3:7]})
860 }
861 if(cur.p[8]==0){ # if beaufort is currently not in the model, propose to add it
862   newparams[8] <- rnorm(1,rjprop.mean.bft,rjprop.sd.bft)
863   new.lik <- log.lik(y = y, x = x, params = newparams, re.obs.int = re.obs.int, re.obs.ph =
864 re.obs.ph)
865   cur.lik <- log.lik(y = y, x = x, params = params, re.obs.int = re.obs.int, re.obs.ph = re.obs.ph)
866   num <- new.lik + sum(log(dunif(newparams[8],prior.params.lo,prior.params.hi))) # add priors
867 for new parameters
868   den <- cur.lik + sum(log(dnorm(newparams[8],rjprop.mean.sp,rjprop.sd.sp))) # add
869 proposal densities for new parameters
870   A<-min(1,exp(num-den))
871   V<-runif(1)
872   ifelse(V<=A,{params[8]<-newparams[8];cur.p[8]<-1},{newparams[8]<-params[8]})
873 }
874 else{ # if beaufort is currently in the model, propose to delete it
875   newparams[8] <- 0
876   new.lik <- log.lik(y = y, x = x, params = newparams, re.obs.int = re.obs.int, re.obs.ph =
877 re.obs.ph)
878   cur.lik <- log.lik(y = y, x = x, params = params, re.obs.int = re.obs.int, re.obs.ph = re.obs.ph)
879   num <- new.lik + sum(log(dnorm(params[8],rjprop.mean.bft,rjprop.sd.bft))) # add proposal
880 densities for current parameters
881   den <- cur.lik + sum(log(dunif(params[8],prior.params.lo,prior.params.hi))) # add priors for
882 current parameters
883   A<-min(1,exp(num-den))
884   V<-runif(1)
885   ifelse(V<=A,{params[8]<-newparams[8];cur.p[8]<-0},{newparams[8]<-params[8]})
886 }
887 # which model did we end up with?
888 cur.mod<-match(sum(cur.p),apply(param.list,1,sum))
889
890 ##### the MH step
891 newparams <- params
892 new.re.obs.int <- re.obs.int
893 new.re.obs.ph <- re.obs.ph
894 # updating the parameters
895 # the first level of species coefficients or beaufort coefficients are always zero, don't need
896 updating
897 for (p in which(cur.p==1)) { # paramters which can be negative
898   u <- rnorm(1,params[p],mhprop.sd.params[p])
899   newparams[p] <- u
900   new.lik <- log.lik(y = y, x = x, params = newparams, re.obs.int = re.obs.int, re.obs.ph =
901 re.obs.ph)
902   cur.lik <- log.lik(y = y, x = x, params = params, re.obs.int = re.obs.int, re.obs.ph = re.obs.ph)

```

```

903     num <- new.lik + log(dunif(newparams[p],prior.params.lo,prior.params.hi))
904     den <- cur.lik + log(dunif( params[p],prior.params.lo,prior.params.hi))
905     A<-min(1,exp(num-den))
906     V<-runif(1)
907     ifelse(V<=A,params[p]<-newparams[p],newparams[p]<-params[p])
908   }
909   for (p in 9:11) {           # st dev cannot be negative
910     u <- rnorm(1,params[p],mhprop.sd.params[p])
911     newparams[p] <- u
912     new.lik <- log.lik(y = y, x = x, params = newparams, re.obs.int = re.obs.int, re.obs.ph =
913 re.obs.ph)
914     cur.lik <- log.lik(y = y, x = x, params =  params, re.obs.int = re.obs.int, re.obs.ph = re.obs.ph)
915     num <- new.lik + log(dunif(newparams[p],prior.sd.lo,prior.sd.hi))
916     den <- cur.lik + log(dunif( params[p],prior.sd.lo,prior.sd.hi))
917     A<-min(1,exp(num-den))
918     V<-runif(1)
919     ifelse(V<=A,params[p]<-newparams[p],newparams[p]<-params[p])
920   }
921
922 # random effects coefficients - no priors on the coefficients
923 for (r in 1:n.obs){
924   new.re.obs.int[r] <- rnorm(1,re.obs.int[r],mhprop.sd.sd.obs.int)
925   num <- log.lik(y = y, x = x, params = params, re.obs.int = new.re.obs.int, re.obs.ph = re.obs.ph)
926   den <- log.lik(y = y, x = x, params = params, re.obs.int =  re.obs.int, re.obs.ph = re.obs.ph)
927   A<-min(1,exp(num-den))
928   V<-runif(1)
929   ifelse(V<=A,re.obs.int[r]<-new.re.obs.int[r],new.re.obs.int[r]<-re.obs.int[r])
930 }
931 for (r in 1:n.obs){
932   new.re.obs.ph[r] <- rnorm(1,re.obs.ph[r],mhprop.sd.sd.obs.ph)
933   num <- log.lik(y = y, x = x, params = params, re.obs.int = re.obs.int, re.obs.ph = new.re.obs.ph)
934   den <- log.lik(y = y, x = x, params = params, re.obs.int = re.obs.int, re.obs.ph =  re.obs.ph)
935   A<-min(1,exp(num-den))
936   V<-runif(1)
937   ifelse(V<=A,re.obs.ph[r]<-new.re.obs.ph[r],new.re.obs.ph[r]<-re.obs.ph[r])
938 }
939
940 # each "n.thin-th" iteration, store the parameter values in matrices
941 if (b %% n.thin < 1) {
942   isave <- isave + 1
943   params.mat[isave,] <- params
944   re.obs.int.mat[isave,] <- re.obs.int
945   re.obs.ph.mat[isave,] <- re.obs.ph
946   model[isave] <- cur.mod

```



```

947 }
948 # display progress
949 Sys.sleep(1)
950 setTkProgressBar(pb, b, label=paste(round(b/n.iter*100),"% completed",sep=""))
951 }          ### end of iteration loop
952 close(pb); date()
953
954 ##### end of RJMCMC sampling
955 #####
956

```

957

958 Appendix 5. Coverage of predicted group sizes

959 For each group size estimate for each observer, we predicted group size using Eq. 1 and
960 sampling the MCMC chains from model 2 as described in Methods. For each of the 2,435
961 observer estimates, we determined if the 95% credibility interval of predicted size included the
962 photo count (our measure of true group size). Coverage of the 95% interval, measured as the
963 fraction of intervals which included the photo count, was 0.955. We note that this procedure was
964 an inverse prediction – that is, although the model fitted y to x , we predicted x given y . We also
965 note that this procedure was not cross-validation, since the model was not refit for each of the
966 2,435 observer estimates. Therefore, since the value being predicted (photo count) was included
967 in the model fitting, coverage was expected to be positively biased. Given the large sample size,
968 however, we believe the positive bias due to the inclusion of a single datum would be small, as
969 indeed it seemed to be. Fig A4 shows observer estimates and posterior distributions of predicted
970 group size plotted against photo count for a selection of the 59 observers.

971

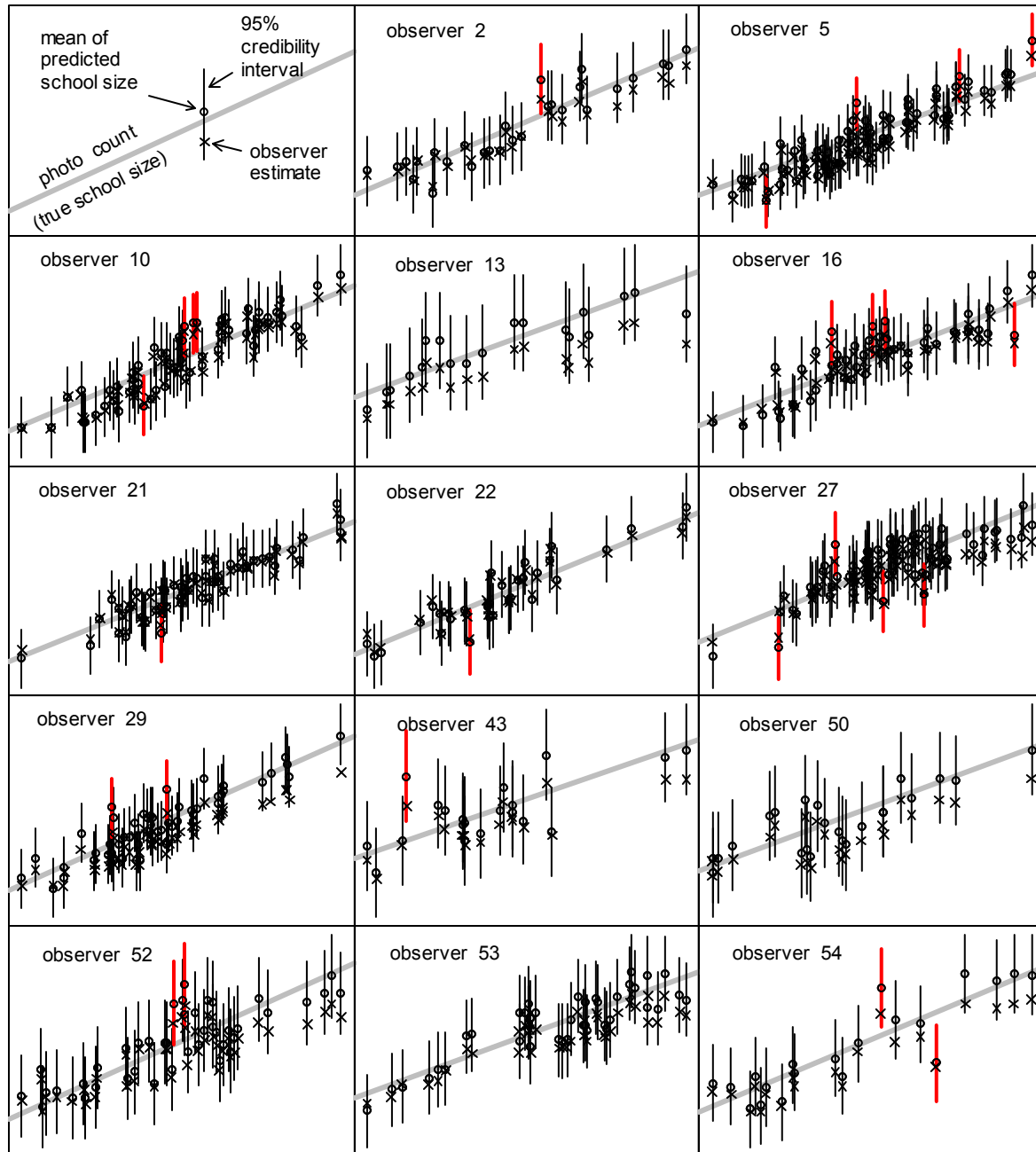


Fig. A4. Observer estimates (x), and group sizes predicted from those estimates, plotted relative to photo count (gray line) for selected observers. Circles are the means and vertical line segments the 95% credibility intervals of predicted group sizes. Cases for which the 95% credibility interval did not include the photo count are shown in red.